Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5248–5251, December 1977
Chemistry

Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins

(protein x-ray structures/hydrophobic interactions/side-chain orientations)

S. RACKOVSKY AND H. A. SCHERAGA*

Department of Chemistry, Cornell University, Ithaca, New York 14853

Contributed by Harold A. Scheraga, September 29, 1977

ABSTRACT **The radial distributions of the C^{α} and side-chain atoms in a sample of 13 native proteins have been examined. It is found that there are substantial differences in behavior between different atoms of the same amino acid. In particular, the C^{α} atoms of polar residues show no particular preference for being far from the center of mass. In light of these results, a new criterion for hydrophobicity and hydrophilicity is proposed—namely, the orientational preference of the side chain. The distribution of this property is shown, and it is suggested that this provides a basis for incorporating hydrophobic interactions into a protein folding algorithm.**

An outstanding difficulty in the calculation of the native conformation of a protein is the multiple-minimum problem. Because of the complexity of the energy surface of a protein as a function of its dihedral angles, energy-minimization programs usually terminate at a local energy minimum rather than at the global minimum. A complete search of the energy surface is impractical because each energy evaluation requires approximately N^2 calculations, N being the number of atoms in the protein.

In order to circumvent these difficulties, approximate folding procedures have been advanced. These are designed to eliminate direct energy calculations in the early stages of the computation, with the approximations being abandoned in the final stages in which energy minimization is carried out (1). These methods utilize various conformational characteristics observed in native proteins in order to approximate the native structure of the molecule under study.

One such property is the polar or nonpolar character of the various amino acid residues. It has usually been assumed that polar residues prefer to be on the outside of the protein, and nonpolar residues on the inside. This phenomenon has been examined in terms of the accessibility of various residues to the solvent (2–6), the proximity of residues to the surface of the protein (7), and the frequency of internal contacts between residues (8, 9). Lee and Richards (2) and others (8, 9) have pointed out that, although the "inside–outside" rule is obeyed in general, exceptions do occur.

Although these studies have contributed to our understanding of protein structure, it is difficult to base an approximate folding procedure on the location of the molecular surface or solvent-accessible atoms. For this purpose, a more useful formulation of hydrophobicity and hydrophilicity is based on the simple notion of distance from the center of mass of the molecule. Polar residues are presumed to prefer to be farther from the center of mass, and nonpolar residues to be nearer. It can be seen readily that, in a given conformation, the determination of the distance of all the atoms from the center of mass requires only $10N$ calculations. This approach has actually been used in a simplified, qualitative form (10).

We are not aware of any quantitative studies of the validity of this formulation of hydrophobicity or hydrophilicity. It was with this objective in mind that the present work was undertaken. We shall see that the distance of a selected atom in a residue from the center of mass does not provide an adequate index of the hydrophobicity or hydrophilicity of the residue, and that different parts of each residue have distinctly different behavior. On the basis of these findings, we shall propose an alternative N-dependent definition of hydrophobicity or hydrophilicity and show that it agrees with traditional, qualitative ideas about this important property.

METHOD

We have studied the distribution of distances of each type of amino acid from the center of mass in a sample of 13 proteins. In order to compare the results from the diverse members of the sample, it is necessary to substitute for actual distance a reduced variable in terms of which the size differences between proteins disappear. To this end, we define the reduced distance r from the center of mass as the actual distance divided by the root-mean-square radius of gyration of the protein. This parameter has a direct physical meaning: a reduced distance greater than 1 means that the atom in question is farther from the center of mass than the average atom in the protein.

It should be noted that there is no direct relationship between reduced distance from the center of mass and internal contacts, solvent accessibility, or presence on the surface of the protein, although there is a general correlation. This is because it is possible for a protein to have a shape (e.g., a long, thin cigar shape) for which a sphere of radius s, centered at the center of mass, extends beyond the actual surface of the molecule in some places.

The proteins comprising the sample were chosen to satisfy three criteria: (*i*) that reliable x-ray coordinates be available (*ii*) that the members of the set not be homologous or closely related, and (*iii*) that the members of the set be mainly single-unit proteins. The last criterion was established because of the possibility that multisubunit proteins might have rather specialized distributions of amino acids in order to bring about the necessary association of subunits.

The protein coordinates were obtained from the Protein Data Bank at the Brookhaven National Laboratory. The following proteins were used: bovine pancreatic trypsin inhibitor, concanavalin A, carboxypeptidase A, flavodoxin, thermolysin, oxidized *Chromatium* high-potential iron protein, staphylococcal nuclease, sea lamprey hemoglobin, hen egg-white ly-

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviation: RDF, radial distribution function.

* To whom reprint requests should be addressed.

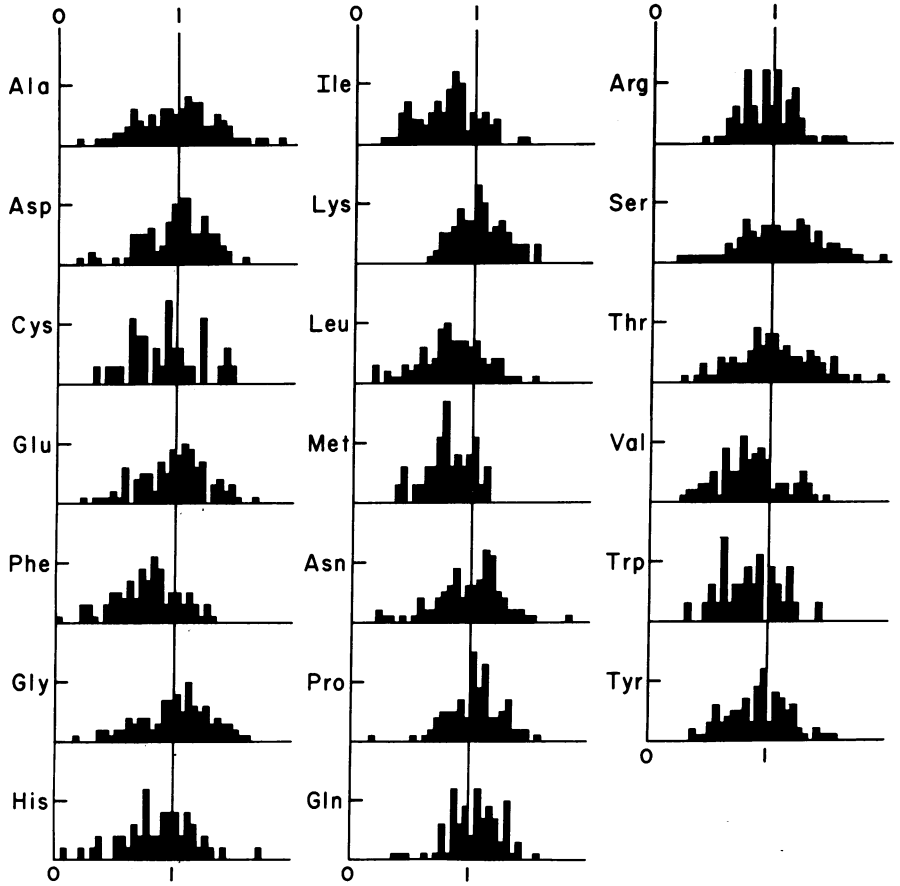

FIG. 1. RDFs for the C^{α} atoms of 20 amino acids. The ordinate represents the fraction of residues occurring in a given r interval. The abscissa is reduced distance r in units of 0.05. We have indicated $r = 1$ and $p = 0.1$ (p, fraction of residues).

sozyme, sperm-whale myoglobin, ribonuclease S, subtilisin BPN′, and rubredoxin.

RESULTS

The radial distribution functions (RDFs) for the C^{α} atoms of the 20 amino acids are shown as histograms in Fig. 1. This figure reveals that most of the residues that traditionally have been considered to be nonpolar exhibit C^{α} RDFs that are centered at some value of $r < 1$. These include Cys, Phe, Ile, Leu, Met, Val, and Trp. The remaining residues, however, exhibit rather *diffuse RDFs centered at* $r \approx 1$. This means that these residues have *about as many* C^{α}*s at* $r < 1$ *as at* $r > 1$. Furthermore, there does not seem to be any relationship between the widths of the distributions and their centers.

Table 1. $\langle r \rangle$ for C^{α} and side-chain atoms

Residue	Side-chain atom	$\langle r \rangle_{C^{\alpha}}$	$\langle r \rangle_{\text{side chain}}$
Ala	C^{β}	0.934	0.941
Asp	$O^{\delta 1}$	0.994	1.071
Cys	S^{γ}	0.900	0.866
Glu	$O^{\epsilon 1}$	0.986	1.100
Phe	C^{ζ}	0.773	0.723
Gly	—	1.015	—
His	$N^{\epsilon 2}$	0.882	0.911
Ile	$C^{\delta 1}$	0.766	0.742
Lys	N^{ζ}	1.040	1.232
Leu	$C^{\delta 1}$	0.825	0.798
Met	C^{ϵ}	0.804	0.781
Asn	C^{γ}	0.986	1.038
Pro	C^{γ}	1.047	1.093
Gln	C^{δ}	1.047	1.150
Arg	$N^{\eta 1}$	0.962	1.112
Ser	O^{γ}	1.056	1.082
Thr	$C^{\gamma 2}$	1.008	1.043
Val	$C^{\gamma 1}$	0.825	0.817
Trp	$C^{\eta 2}$	0.848	0.867
Tyr	O^{η}	0.931	1.050

Table 2. Number of occurrences of orientations

Residue	Number with $0 < \theta < \pi/2$	Number with $\pi/2 < \theta < \pi$
Ala	113	97
Asp	93	35
Cys	12	24
Glu	72	30
Phe	23	53
His	24	28
Ile	43	75
Lys	113	29
Leu	46	90
Met	10	25
Asn	77	39
Pro	55	27
Gln	58	15
Arg	43	25
Ser	114	71
Thr	83	56
Val	55	93
Trp	15	20
Tyr	62	36

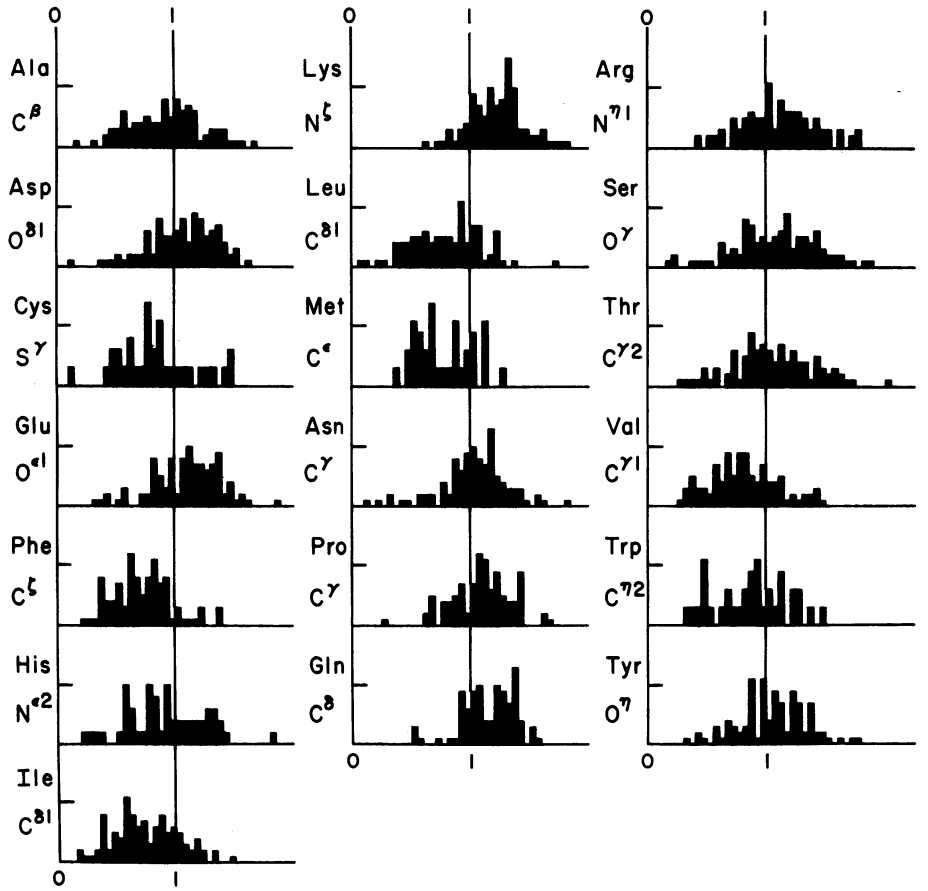

FIG. 2. RDFs for the indicated side-chain atoms. Conventions as in Fig. 1.

A useful parameter for characterizing the RDFs is the average reduced distance $\langle r \rangle$ defined by

$$\langle r \rangle = \sum_i p_i r_i$$

in which p_i is the fraction of residues of the given type in interval i, and r_i, the reduced radius at the center of interval i, is given by $r_i = (0.05)(i - 1) + 0.025$. The values of $\langle r \rangle$ for the C^{α} atoms of the 20 amino acids are listed in column 3 of Table 1. There are 13 residues for which $0.9 \leq \langle r \rangle \leq 1.056$. This is a quantitative representation of our previous observation that most of the residues that are not nonpolar have C^{α} RDFs centered at $r \approx 1$.

In view of these results, it is of interest to examine RDFs for appropriate atoms located near the ends of side chains. These are exhibited in Fig. 2 and in column 4 of Table 1. We see that side-chain atoms exhibit less neutral behavior; those residues that are traditionally regarded as polar frequently have side-chain atom RDFs that are centered at $r > 1$.

Clearly, hydrophilicity is a more complex phenomenon than generally has been thought, in the sense that different parts of a residue exhibit the phenomenon in markedly differing degrees. For example, there is a large difference between the RDFs for the C^{α} and N^{ζ} of lysine. This suggests that the variable best suited for defining the hydrophilicity or hydrophobicity of a residue is the *orientation of the side chain*, which reflects the information contained in both sets of RDFs. Therefore, we studied the distribution of θ, the angle between the center-of-mass-to-C^{α} vector and the C^{α}-to-side-chain-atom vector (Fig.

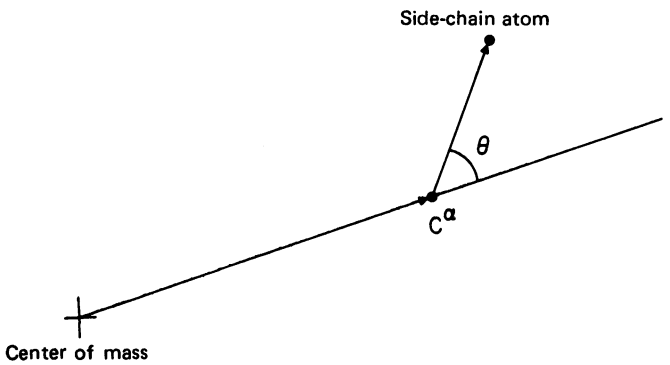

FIG. 3. Definition of θ.

Table 3. Amino acids by orientational preference*

$N_{>\pi/2} < N_{<\pi/2}$	$N_{>\pi/2} > N_{<\pi/2}$
Ala	Cys
Asp	Phe
Glu	His
Lys	Ile
Asn	Leu
Pro	Met
Gln	Val
Arg	Trp
Ser	
Thr	
Tyr	

* Because glycine has no side chain, it has no orientational preference.

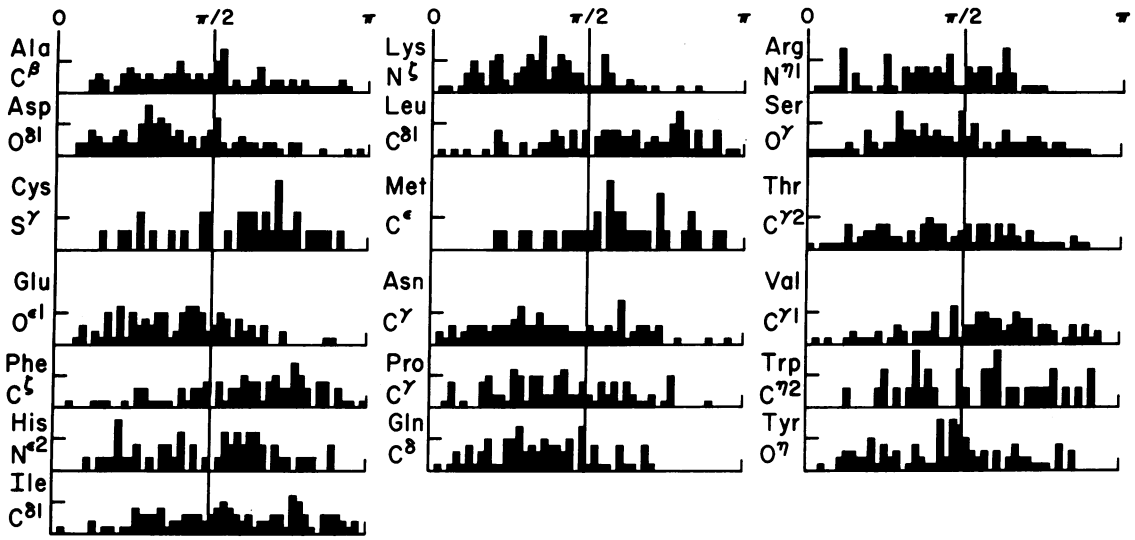

FIG. 4. Distribution functions for θ. The ordinate represents the fraction of residues occurring in a given θ interval. The abscissa is θ in units of 3.6°. We have indicated $\theta = 0, \pi/2$, and π on the abscissa and 0.1 on the ordinate.

3). This, too, is an N-dependent property for a given conformation. These distributions are presented in Fig. 4.

In these distributions, the residues cited above as being traditionally regarded as nonpolar show a predominance of values of $\theta > \pi/2$. In polar residues, $\theta < \pi/2$ predominates. These data are summarized in Tables 2 and 3.

CONCLUSIONS

Neither the C^{α} RDF nor the side-chain RDF is a reliable indicator of the placement of a residue in a protein. In order to incorporate the influence of hydrophobic interactions properly, it is necessary to specify the *orientation* of the side chain. This has implications for protein folding schemes, because orienting the side chain will automatically constrain the backbone dihedral angles ϕ and ψ. It also suggests that, in any simple (approximate) model of a protein, some representation must be made of both the C^{α} atom and the side chain; single-bead models are of limited usefulness.

Work is now continuing directed toward incorporating these results into a protein folding algorithm.

We thank Dr. G. Némethy for helpful discussions. This work was supported by research grants from the National Science Foundation (PCM75-08691) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312). S.R. is a National Institutes of Health Postdoctoral Fellow, 1977–1978.

1. Nemethy, G. & Scheraga, H. A. (1977) *Q. Rev. Biochem. Biophys.* **10**, 239–352.
2. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55**, 379–400.
3. Shrake, A. & Rupley, J. A. (1973) *J. Mol. Biol.* **79**, 351–371.
4. Chothia, C. (1974) *Nature* **248**, 338–339.
5. Chothia, C. (1975) *Nature* **254**, 304–308.
6. Chothia, C. (1976) *J. Mol. Biol.* **105**, 1–14.
7. Wertz, D. H. & Scheraga, H. A. (1978) *Macromolecules*, in press.
8. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9**, 945–950.
9. Tanaka, S. & Scheraga, H. A. (1977) *Macromolecules* **10**, 291–304.
10. Kuntz, I. D., Crippen, G. M., Kollman, P. A. & Kimelman, D. (1976) *J. Mol. Biol.* **106**, 983–994.